The Tutorbot Corpus – A Corpus for Studying Tutoring Behaviour in Multiparty Face-to-Face Spoken Dialogue

Maria Koutsombogera[(2)], Samer Al Moubayed[(1)], Bajibabu Bollepalli[(1)], Ahmed Hussen Abdelaziz[(3)], Martin Johansson[(1)], José David Aguas Lopes[(4)], Jekaterina Novikova[(5)], Catharine Oertel[(1)], Kalin Stefanov[(1)], Gül Varol [(6)]

[1] KTH Speech, Music and Hearing, Sweden
[2] Institute for Language and Speech Processing – Athena R.C., Greece
[3] Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany
[4] Spoken Language Systems Laboratory, INESC ID, Portugal
[5] Department of Computer Science, University of Bath, UK
[6] Department of Computer Engineering, Bogazici University, Turkey
mkouts@ilsp.athena-innovation.gr, sameram@kth.se, bajibabu@kth.se, ahmed.hussenabdelaziz@rub.de, vhmj@kth.se, zedavid@l2f.inesc-id.pt, j.novikova@bath.ac.uk, catha@kth.se, kalins@kth.se, gul.varol@boun.edu.tr

Abstract

This paper describes a novel experimental setup exploiting state-of-the-art capture equipment to collect a multimodally rich game-solving collaborative multiparty dialogue corpus. The corpus is targeted and designed towards the development of a dialogue system platform to explore verbal and nonverbal tutoring strategies in multiparty spoken interactions. The dialogue task is centered on two participants involved in a dialogue aiming to solve a card-ordering game. The participants were paired into teams based on their degree of extraversion as resulted from a personality test. With the participants sits a tutor that helps them perform the task, organizes and balances their interaction and whose behavior was assessed by the participants after each interaction. Different multimodal signals captured and auto-synchronized by different audio-visual capture technologies, together with manual annotations of the tutor's behavior constitute the Tutorbot corpus. This corpus is exploited to build a situated model of the interaction based on the participants' temporally-changing state of attention, their conversational engagement and verbal dominance, and their correlation with the verbal and visual feedback and conversation regulatory actions generated by the tutor.

Keywords: Multimodal corpus, Multiparty Interaction, Tutor.

1. Introduction

Research in the last years has been moving towards analyzing multiparty, multimodal conversations with the aim to understand and define the structure and strategies with which interlocutors regulate the interaction, and keep their conversations rich, fluent and successful. Building spoken dialogue systems has the potential of not only providing a hands-free interface for information input and output, but even more importantly, the ability of using speech technology to provide a human-like interface that can understand and communicate all the subtle non-verbal signals that accompany the stream of sounds and provide significant information about the interpretation of the state of the user and his actions. These signals become even more central in scenarios where affective and social skills are essential for the success of the interaction (such as learning, collaborative task solving, games and commerce (Nass, Steuer & Tauber, 1994; Cohen, 1992; Cohen & Oviatt, 1995)).

In this paper, we present the design and the capture of a novel multiparty dialogue setup and corpus collected to explore and build a collaborative tutoring agent engaged in a game-solving task with two human-interlocutors. The work presented here is part of a larger project carried out at the eNTERFACE'13[1] Multimodal Interfaces Workshop. Our work addresses social and interactional skills required by the dialogue system to control the interaction flow as well as to boost and balance the engagement of the participants in the task they are involved in, while at the same time mitigating dominant behavior and encouraging less talkative interlocutors to equally participate in the interaction.

While developing our setup, we target the ability to move directly from the models learnt from the annotations and analysis of the corpus into an implementation of multiparty multimodal dialogue system using the robot head Furhat (Al Moubayed et al., 2012) and the newly developed IrisTK dialogue platform (Skantze & Al Moubayed, 2012), both developed and utilized in multimodal multiparty embodied spoken dialogue systems.

2. Experimental design and setup

The task implemented was a shortened version of a "*NASA Exercise: Survival on the Moon*". Participants have to imagine that they are members of a space crew that has landed 200 miles away from their mother ship, thus they have to walk this distance carrying items critical for their survival. The participants were presented with 6 cards depicting the objects of interest (oxygen, food, water, first aid kit, stellar map, magnetic compass). They were then asked to discuss each of the six cards and rank them in terms of their importance - and in line with the correct NASA's ranking - by collaborating with each other to reach a consensus. The task was presented and supervised by a human tutor.

The experimental setup consisted of a tutor and 2

[1] http://eventos.fct.unl.pt/enterface13/home

participants sitting around a table at approximately equal distance from one another, composing an equilateral triangle (Figure 1). At the beginning of each recording, the tutor presented the task by explaining the game and initiated the interaction by asking the participants to pick a card and flip it to see the object. After some time, when an agreement had been reached, the tutor instructed the participants to return the card and move on to the next. After all 6 cards had been discussed, the tutor requested from the participants to give their final decision on the correct order.

Eight sessions of recordings were carried out, each consisting of the same tutor and a different pair of participants. Participants were asked to fill in a Big Five Personality Test (Digman, 1990) and were then grouped into pairs based on their results, with a focus on the percentile ranks of the *extraversion* dimension. The extraversion trait was favored since it is linked to the engagement of a subject in a task, his/her involvement in the social world, and most importantly, the degree of enthusiasm and enjoyment in interaction with the co-locutor. In this respect, it is related to cues that are important for this experiment, namely the conversational engagement and dominance.

The pairs were formed in a way that one of two team members scored high on extraversion and the other one scored low. The average difference between participants on the extraversion dimension was 28 points. Differences in the percentiles among pairs may vary, as shown in the last column of table 1; given the short duration of the project, the recruitment of subjects dynamically depended on their availability and kept on even as the first recordings were being carried out.

The tutor's interactional behavior followed 2 experimental conditions, each applied to 4 sessions: (a) the *good* tutor, i.e. actively dedicated to the participants' success, supportive, friendly and managing turn-taking so that both participants engage in the discussion and the decision making. (b) The *neutral* tutor, on the other hand, delivered material in a clear and concise manner but didn't need to make the communication either interesting or enjoyable or influence the conversation flow. Prior to the recordings, instructions about the two different behavior setups (good - neutral) were given to the person that would act as the tutor, and a couple of sessions were run to test the relevant behavior instantiations.

After the recordings, we tested the participants' ability to recognize the tutor's behavior by having them fill in a Tutor Assessment Questionnaire (cf. Appendix, Figure 4) consisting of 24 pairs of contrasting characteristics that may apply to the tutor, based on similar work on user experience (Laugwitz et al. 2008). The numbers between the characteristics represent gradations between opposites. A seven-stage scale is used for gradation in order to reduce the well-known central tendency bias for such types of items.

The consistency of each group was measured by the Cronbachs Alpha-Coefficient (Cronbach, 1951), varying between 0.72 and 0.95 and indicating that the given groups of items in the questionnaire were consistent and that participants interpreted the items in an expected way in the given task.

3. Technical setup and capture equipment

Kinect[2] cameras were installed in front of each participant for skeletal and face tracking. Collected Kinect data was used for extracting the direction of participants' head rotation and posture. Two more high-definition video cameras were installed in front of and behind the tutor to capture the tutor's behavior as well as the entire scene. To capture the spoken input in a multiparty setup, the Microcone®[3] microphone array was used, allowing for automatic speaker diarization.

The six rectangular cards were placed on the table and were recorded by one high-resolution video camera aiming to track them in real-time to significantly help the dialogue system distinguish among them, even in the case where an ASR module would fail to understand the content of the speech signal. The setup is presented in Figure 1. Figure 2 and 3 show the input from the Kinect camera and the Microcone® respectively.

The recording setup is controlled to a certain extent to maximize the desired output to be acquired. Furthermore, the sessions subjects were all participants of eNTERFACE'13, i.e. students and young researchers who are familiar with the use of similar equipment employed for research tasks.

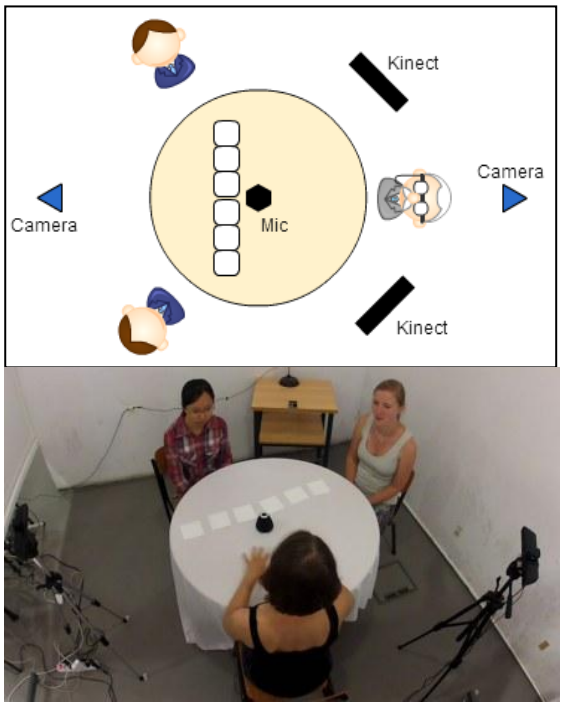

Figure 1: Experimental setup

[2] http://www.microsoft.com/en-us/kinectforwindows/
[3] http://www.dev-audio.com/products/microcone/

Figure 2: An overlay of the head pose and skeleton tracking by Kinect

Figure 3: The Microcone® microphone array (left) and a visualization of an activity of the different microphones (right)

4. Corpus Description & Annotation Process

Eight dialogues were collected, balanced over extroversion in personality of the 2 participants and the tutoring condition, resulting in a final corpus of approximately 82 minutes duration. Table 1 presents the eight different sessions, the gender of the participants in each session, the duration, the tutor condition and the differences of the participants in the extraversion percentiles they scored in their personality tests.

Session	Participants' Gender	Duration	Tutor scenario	Extraversion diff.
1	M-M	13.28	Good	37
2	M-M	15.44	Good	44
3	F-F	07.08	Good	5
4	M-M	10.43	Good	21
5	M-F	07.15	Neutral	9
6	M-M	09.39	Neutral	74
7	M-M	08.07	Neutral	13
8	M-F	10.05	Neutral	21

Table 1: The Tutorbot corpus description.

4.1 Manual Annotation

The data collection was manually annotated with regards to the multimodal conversational behavior of the tutor. The goal being to analyze the multimodal strategies employed to manage the conversation, the annotation was focused on describing forms and functions of the related verbal and non-verbal signals. Multimodal interaction in both its two-party and multi-party dimensions is substantially related to the functions of feedback and turn-taking management, mechanisms that have been thoroughly studied and documented in robust theoretical frameworks (Sacks, Schegloff & Jefferson, 1974; Allwood et al., 1993), including multimodal perspectives (Duncan, 1972) and multiparty turn-taking in dialog systems (Bohus & Horvitz, 2010 & 2011).

The annotation of the recorded video sessions was performed in ELAN [4] (Wittenburg et al., 2006). An annotation scheme was employed to cater for all the features that need to be represented for the task at hand. It is heavily based on widely-used labeling sets designed for annotating multimodal interaction (Allwood et al., 2007; Bunt et al., 2010), and was tailored to the needs of the task. Specifically, the goal of the annotation was to account for multimodal behavior including verbal and nonverbal signals as well as conversational structures and functions expressed in a multimodal manner.

The speech activity of the tutor was transcribed to export utterances that a dialogue system would use to manage the interaction. Speech activity was next attributed a label of a dialogue act describing the communicative action which the tutor performs. Questions of the tutor are labelled depending on the addressee, i.e. *speaker*, *listener*, or *both participants*. In this experimental setup, the cues that the tutor provides to aid the participants have a central role (i.e. *hint*). *Instructions* and *requests* to the participants are given throughout the discussion. The scheme also caters for *answers* that the tutor gives to the participants or instances of *agreement* or *disagreement* with them

Labels related to feedback and turn management were horizontally attributed to cover both functions of verbal and non-verbal attestations (head movements, facial expressions, facial gestures), cues that are considered of high importance to the regulation of the interaction (i.e. the way the tutor takes, holds or assigns the turn) as well as to feedback expression. Distinct values signal the normal transition of turns (*take*, *accept*, *complete*, *offer*), but also cases of interruptions and overlapping talk (i.e. *grab*, *yield*, *hold*).

The set of feedback labels describe whether the tutor gives or elicits signs of perception and understanding, agreement or disagreement with what the participants say. They apply to backchannels or the expression of evaluations as well as to head movements and facial expressions (e.g. nodding, smiling).

Non-verbal signals were identified on the time axis and then marked first according to their form and subsequently to their function, i.e. whether they have a feedback or a turn management purposes. Since the corpus development targets the implementation of a dialogue system with the Furhat robot head, the annotation of non-verbal expressions targets the head area. Specifically, annotated non-verbal expressions include general facial expressions (e.g. smile, scowl), head movements (e.g. nod, turn, shake), gaze (left, right, towards cards), eyes (e.g. wide open), eyebrows (raising, frowning) and mouth movements (e. g. open, protruded

[4] ELAN (http://www.lat-mpi.eu/tools/elan/)

lips). These signals are indicative of the state of mind of the tutor, the level of perception of the discussion, the stance towards the participants (agreement, disagreement, encouragement), the expression of feedback and the regulation of the interaction.
Finally, the annotation scheme includes a layer indicating each card id and the time interval within which it is being discussed.Layers and values of the annotation scheme can be found in the Appendix, Table 3.

4.2 Capturing audiovisual signals

Automatic measurements of the participants' behavior were conducted through capturing by the respective equipment with the goal of building related modules that can be exploited by the dialogue system, integrated with the manual annotations of the tutor's behavior.
Audio data were captured with the Microcone™ microphone array, which also provides a stream of the microphone activation status for the audio channels corresponding to each session participant, allowing for the detection of multiple speakers, and hence overlaps and speakers locations. This stream was exploited to implement a voice activity detector module for the dialogue system.
Data captured by the two kinects include the head pose and skeletal movement of the experiment participants, as well as parameters describing their facial expressions. Head position and orientation is exploited to infer the focus of attention of the participants and to build the visual attention module.

5. Data analysis and conversational management strategies

This section presents preliminary results from the analysis of the conversational behaviors the tutor exhibited in the corpus. Data analysis was carried out to investigate differences between the *good* and the *neutral* tutor condition. To this end we used low-level cues such as voice activity, gaze and facial movements as well as their timing. Furthermore, differences in the tutor's dialogue acts and turn management behavior (in terms of frequency and different values employed) were investigated.
Our results indicate that turn management behaviour conveys essential and richer information compared to the dialogue acts types used, i.e. the timing and the conversation managing action of what is said matters more than the actual content per se. For example, the number of turn offers as well as turn accepts is significantly higher in the *good* tutor condition than in the *neutral* one (42 vs. 8 and 33 vs. 13 respectively). We also hypothesized that: (a) the number of dialogue acts such as hints or instructions the tutor gives will be higher for the *good* tutor condition than for the *neutral* one and (b) the tutor will employ more turn management features in the *good* tutor condition than the *neutral* tutor condition. Concerning hypothesis a) we found a difference in the number of hints between *good* and *neutral* tutor condition (29 vs. 27 hints) and concerning hypothesis b) we also found that the number of turn grabs is higher in the *good* than in the *neutral* tutor condition (13 vs. 8 turn grabs).
In this first analysis of results we attempt to capture any overall differences between the good and neutral tutor paradigm, under the assumption that the human tutor was, at least to an acceptable degree, consistent (an assumption supported by the assessment tests results) in the behavior shown in the two experimental settings. A human tutor is not guaranteed to be consistent, yet post-assessment tests are verifying our hypotheses with regards to the experimental design to a large extent. A sample of statistics calculated on features is presented in Table 2 below.

Feature	Good tutor	Neutral tutor
Avg. time of all conversations	11.76 (3.69)min	8.77 (1.3)min
Avg. time on each card	1.29 (0.66)min	0.85 (0.49) min
No. of hints in all conversations	7.25 (2.06)	6.50 (1.91)
No. of agreements	2.00 (1.82)	2.50 (1.91)
No. of disagreements	2.00	1.00
No. of instruction/request	4.25 (0.95)	3.25 (0.5)
No. of turn grabs	3.25 (2.06)	2.25 (2.06)
No. of turn offers	10.50 (8.38)	4.50 (5.74)

Table 2: Statistics of features presented in the following order: mean (standard deviation).

6. Discussion and conclusions

In this paper, we presented an experimental setup for a multiparty dialogue. Such dialogue design can be employed in human-robot spoken dialogue systems able to monitor and regulate the interaction of the interlocutors. The recording of the corpus included devices capturing multimodal signals that would provide real-time input to a dialogue system, while the behavior of the tutor was manually annotated and analyzed. The current analysis shows inherent differences between an active (*good*) and a passive (*neutral*) tutor. We intend to exploit this corpus and findings to further develop a dialogue system able to directly use the data inputted from the capture devices and to exhibit behaviors similar to the tutor's actions that are annotated.
Our design decisions targeted in making this transition step possible: the capture setup used can be directly used in a human-robot interaction; the task design and the use of cards as visual indicators of information lower the need for robust ASR requiring the constraint of natural interaction, currently not available; finally, the use of personality tests and tutoring behaviors also allow for the quantification of different behavioral models (frequency, type, and timing of actions such as dialogue acts and gestures).

7. Acknowledgements

The authors would like to thank the LREC reviewing committee for their valuable feedback, and the eNTERFACE participants for taking part in the recording sessions.

8. References

Al Moubayed, S., Beskow, J., Skantze, G., and Granström, B. (2012). Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction. In Esposito, A., Esposito, A., Vinciarelli, A., Hoffmann, R., & C. Müller, V. (Eds.), *Cognitive Behavioural Systems, Lecture Notes in Computer Science 7403*, Springer, pp. 114--130.

Allwood, J., Nivre, J., and Ahlsén, E. (1993). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1), pp. 1--29.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena. Multimodal Corpora for Modeling Human Multimodal Behaviour. Journal on Language Resources and Evaluation, 41(3-4), pp. 273--287.

Bohus, D., Horvitz, E. (2010). Facilitating multiparty dialog with gaze, gesture, and speech. In Proceedings of ICMI'10, New York: ACM Press.

Bohus, D., Horvitz, E. (2011) - Decisions about Turns in Multiparty Conversation: From Perception to Action. Proc. of ICMI-2011, New York: ACM Press.

Bunt, H. Alexandersson, J., Carletta, J., Choe, J.-W., Fang, A.C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., and Traum, D.R. (2010). Towards an ISO Standard for Dialogue Act Annotation. In *Proceedings of LREC 2010*, ELRA, pp. 2548--2555.

Cohen, P. (1992). The role of natural language in a multimodal interface. In *Proceedings of the 5th annual ACM symposium on User interface software and technology*, ACM, pp. 143--149.

Cohen, P. & Oviatt, S. (1995). The role of voice input for human–machine communication. In *Proceedings of the National Academy of Sciences*, 92 (22), pp. 9921--9927.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16 (3), pp. 297--334.

Digman, J.M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology* 41, pp. 417--440.

Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversation. *Journal of Personality and Social Psychology*, 23, pp. 283--292.

Kinect for Windows (n.d.). Retrieved March 21, 2014, from http://www.microsoft.com/en-us/kinectforwindows/

Laugwitz, B., Held, T., and Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. *HCI and Usability for Education and Work*, Berlin, Heidelberg: Springer, pp. 63--76.

Microcone: The Intelligent Microphone Array for Groups (n.d.). Retrieved March 21, 2014, from http://www.dev-audio.com/products/microcone/

Nass, C., Steuer, J. and Tauber, E. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York: ACM Press, pp. 72--78.

Sacks, H., Schegloff. E., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language*, 50, pp. 696--735.

Skantze, G., Al Moubayed, S. (2012). IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of ICMI'12* New York: ACM Press, pp. 69--76.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of LREC 2006)*, ELRA, pp. 1556--1559.

Appendix

Annotation layers	Values
Speech_activity	Free text
Dialogue acts	Question_speaker, Question_listener, Question_both, Hint, Answer, Instruction/Request, Agreement, Disagreement
Turn management	Take, Accept, Grab, Offer, Complete, Yield, Hold, Backchannel
Feedback	Perception/Understanding (Give-Elicit) Accept (Give-Elicit), Non-accept (Give-Elicit)
Verbal_feedback	Free text
Face_general	Smile, Laugh, Scowl
Head_movement	Nod(s), Shake, Jerk, Tilt, Turn, Forward, Backward
Gaze	Attention_Person_Right, Attention_Person_Left, Attention_Object, Glance
Eyes	Wide_open, Semi-closed, Wink, Blink
Eyebrows	Raise, Frown
Mouth	Open, Closed
Functions_[Face, Head, Gaze, Eyes, Eyebrows, Mouth]	Feedback, Turn Management
Cards	Card id

Table 3: The annotation scheme employed for the manual analysis of the tutor conversational behavior.

Tutor Assessment Questionnaire

For the assessment of the tutor, please fill out the following questionnaire. The questionnaire consists of pairs of contrasting characteristics that may apply to the tutor. The numbers between the characteristics represent gradations between the opposites. You can express your agreement with the characteristic by ticking the number that most closely reflects your impression. Sometimes you may not be completely sure about your agreement with a particular attribute or you may find that the attribute does not apply completely to the particular tutor. Nevertheless, please tick a number in every line.

Please decide spontaneously. Don't think too long about your decision to make sure that you convey your original impression.

It is your personal opinion that counts. Please remember: there is no wrong or right answer!

Name/Nickname (for identification): _____

I think that the tutor is...

		1	2	3	4	5	6	7	
1	annoying								enjoyable
2	unfriendly								friendly
3	bad								good
4	unpleasant								pleasant
5	does not meet expectations								meets expectations
6	socially obstructive								socially supportive
7	unpredictable								predictable
8	slow								fast
9	inefficient								efficient
10	cluttered								organized
11	confusing								clear
12	complicated								easy
13	not understandable								understandable
14	acts as if he/she does not know what to do								acts as if he/she does know what to do
15	doesn't care to hold our attention								holds our attention
16	doesn't give us feedback on the quality of our work								gives us feedback on the quality of our work
17	doesn't help us with our task								helps us with our task
18	doesn't realise when we fail to agree								realises when we fail to agree
19	hesitant								never hesitates
20	if I have something to say, the tutor won't listen								if I have something to say, the tutor will listen
21	impatient								patient
22	not consistent in his/her behavior								consistent in his/her behavior
23	not motivating								motivating
24	passive								active

Figure 4: The Tutor Assessment Questionnaire